Information Retrieval and Search Engines

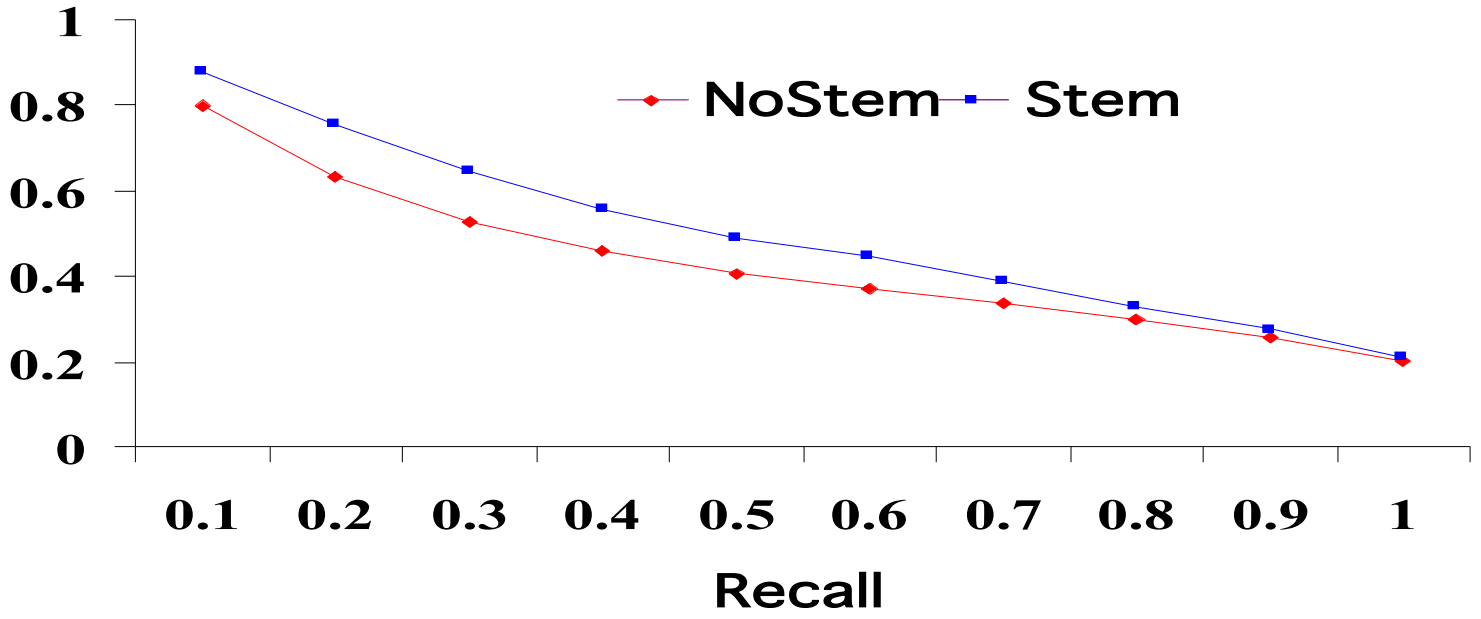
Review
and Pseudo-Final

Week 1 – IR and the Web

- IR defined
  - Data Issues
    - Distributed, Volatile, Large, Unstructured, Redundant, Questionable Quality, Heterogeneous
  - Zipf's Law
    - Fan in, Fan out, Word use, Page length, Number of hits
  - Size of Web
    - At least 24 billion pages
    - Hidden web

Week 2 -- Evaluation

- Relevancy
  - Subjective, Situational, Cognitive, Dynamic
- Recall & Precision
- Parts of a Search Engine
  - Crawler, Indexer, Rertiever



Week 3 – Text & Vector Spaces

- Zipf's Law, Heap's Law, Luhn's Idea
- Data Preparation
  - Tokenization, stemming, stop words, indexing
- Boolean Models
  - Simple
  - Non-intuitive, unordered,
- Vector Space
  - Bag of words, cosine distance, partial matching
  - Missing semantics, missing syntax, term independence assumption, ...

Week 4/6 – More VS & Probabilistic

- TF-IDF weights
- Probability
  - Prior
  - Conditional (or Posterior)
  - Bayes Rule -- P(A|B) = (P(B|A)/P(B)) P(A)
  - Naïve Bayes Classifier
  - $$\sum_{y \in present, absent} \sum_{x \in (yes, no)} P(word = y, isSpam = x) * \log\left(\frac{(P(word = y, isSpam = x))}{(P(word = y) * P(isSpam = x))}\right)$$

Week 7 -- Google

- PageRank
  - Based only on “authority”
  - $$R(p) = c \sum_{q:q \to p} \frac{R(q)}{N_q}$$
  - Rank sinks and rank sources
    - The “god” node
  - Question: Suppose you add a link from page A to page B. What would you expect to happen? Why?
    - Suppose that no change occurred to rank of B. Why?
- PageRank based spidering.

Week 8 – H&A & LSI

- Hubs and Authorities
  - Hub vs authority
    - Local calculation vs global for PageRank
- Latent semantic Indexing – why not in general use?

Pros and Cons for LSI

- **Pro:**
  - Can improve retrieval and overcome "vocabulary match" problem
  - Gives lower-dimensional vectors
    - Nice for machine learning algorithms that cannot handle high dimensional spaces
    - Each dimension more "semantic" (?)
- **Contra:**
  - New vectors are dense
    - We do not necessarily save memory
    - Inverted index does not work for dense vectors => less efficient retrieval
  - Words with several meanings confounded by LSI
  - Expensive to compute, but can be done offline

Week 9 -- Clustering

- Ugly Duckling Theorem
- Distance Metrics
  - L1 vs L-infinity
- HAC vs Divisive
  - Hard vs soft
  - Single vs complete link
- K-means
  - Buckshot, Fractionization

EM Algorithm

- Iterative method for learning probabilistic categorization model from unsupervised data.
- Initially assume random assignment of examples to categories.
- Learn an initial probabilistic model by estimating model parameters from this randomly labeled data.
- Iterate following two steps until convergence:
  - Expectation (E-step): Compute P(c_i | E) for each example given the current model, and probabilistically re-label the examples based on these posterior probability estimates.
  - Maximization (M-step): Re-estimate the model parameters, , from the probabilistically re-labeled data.

Week 12 – Collaborative Filtering

- Approach
  - Maintain a database of many users' ratings
  - find other similar users
  - Recommend items rated highly by these similar users
- Problems
  - Cold-start, Sparsity, First-Rater, Popularity-Bias
- Evaluation
  - Task dependency of criterion, leakage
-

Week 13- RF

- Relevance Feedback
  - Rocchio
    - Near misses
    - Single Hyperplane / Hypersphere
- The Routing & Filtering task
- Adaboost
  - Why Adaboost is equivalent to Rocchio